# MINING FORUMS FOR SOLUTIONS TO QUESTIONS

## BACKGROUND OF THE INVENTION

**[0001]** There is a wealth of information embedded inside the text of user posts within threaded online discussions, such as forums and bulletin boards. A challenge, however, is that the information is scattered across pages, users, and even sites. Furthermore, the information is unstructured and often extremely difficult to follow. Moreover, the information within such user posts often suffers from the problem of unreliable quality. A statement made within a post may be incorrect or off-topic. Because of these issues, traditional automated analysis of discussion threads work with the meta-data and structural information, which can be used for discovering topic heat maps, finding contentious discussions based on thread length, or determining power users within a particular forum.

**[0002]** A great number of discussion forums focus on providing expertise and help to a community of interest. Discussion threads on these forums originate with the posting of a question or problem to be solved, and replies to the original post take the form of answers to the question. The syntactic structure and information content of individual sentences that occur within a dialogue is different from that found in monologue sentences that appear within a narrative. This makes it difficult to analyze threaded discussion posts using parsing and other natural language processing (NLP) techniques developed for written monologue.

## SUMMARY

**[0003]** An approach is provided for mining threaded online discussions. In the approach, performed by an information handling system, a natural language processing (NLP) analysis is performed on threaded discussions pertaining to a given topic. The analysis is performed across multiple web sites with each of the web sites including one or more threaded discussions. The analysis results in harvested discussions pertaining to the topic. The harvested discussions are correlated and a question is identified from the harvested discussions. A set of candidate answers is also identified from the harvested discussions, with one of the candidate answers being selected as the most likely answer to the identified question.

**[0004]** The foregoing is a summary and thus contains, by necessity, simplifications, generalizations, and omissions of detail; consequently, those skilled in the art will appreciate that the summary is illustrative only and is not intended to be in any way limiting. Other aspects, inventive features, and advantages of the present invention, as defined solely by the claims, will become apparent in the non-limiting detailed description set forth below.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0005]** The present invention may be better understood, and its numerous objects, features, and advantages made apparent to those skilled in the art by referencing the accompanying drawings, wherein:

**[0006]** FIG. **1** depicts a network environment that includes a knowledge manager that utilizes a knowledge base;

**[0007]** FIG. **2** is a block diagram of a processor and components of an information handling system such as those shown in FIG. **1**;

**[0008]** FIG. **3** is a component diagram depicting the various components in mining threaded online discussions;

**[0009]** FIG. **4** is a depiction of a flowchart showing the logic used in site discovery of threaded online discussions and harvesting content from such discussions;

**[0010]** FIG. **5** is a depiction of a flowchart showing the logic used to classify discussion posts and update a corpus utilized by a deep question answering system;

**[0011]** FIG. **6** is a depiction of a flowchart showing the logic performed by the question answering pipeline;

**[0012]** FIG. **7** is a depiction of a flowchart showing the logic used by the system to generate candidate answers;

**[0013]** FIG. **8** is a depiction of a flowchart showing the logic performed by the system to prune a contribution tree of unneeded or superfluous contribution posts; and

**[0014]** FIG. **9** is a depiction of a flowchart showing the logic performed to add candidate answers to the set of answers identified for consideration as the most likely answer to the identified question.

## DETAILED DESCRIPTION

**[0015]** As will be appreciated by one skilled in the art, aspects of the present invention may be embodied as a system, method or computer program product. Accordingly, aspects of the present invention may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "module" or "system." Furthermore, aspects of the present invention may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

**[0016]** Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device.

**[0017]** A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport